Real-time virtual reference service based on applicable artificial intelligence technologies:
The début of the robot Xiaotu at Tsinghua University Library

Fei YAO¹, Lei Ji², Chengyu ZHANG¹* & Wu CHEN¹

¹ Tsinghua University Library, Beijing 100084, China
² Agricultural Bank of China, Main Branch, Beijing 100005, China

Abstract The adoption of applicable artificial intelligence technologies to library real-time virtual reference services is an innovative experimentation in one of the key areas of library services. Based on the open source software Artificial Linguistic Internet Computer Entity (A.L.I.C.E.) and a combined application of several other relevant supporting technologies for facilitating the use of the current existing library resources, Tsinghua University Library has recently developed a real-time smart talking robot, named Xiaotu, for the enhancement of its various service functions, such as reference services, book searching, Baidu Baike searching, self-directed learning, etc. The operation of Xiaotu is programmed into Renren website (a social networking website), which adds significantly an innovative feature to the modus operandi of the real-time virtual reference service at Tsinghua University Library.

Keywords Virtual reference service, Smart talking robot, Self-learning, A.L.I.C.E.

1 Introduction

Virtual reference service has been a novel and growing part of library operation since its birth in mid-1990s. Its service mode has evolved from the initial stage of a single-dimensional mode of service delivery such as e-mail and web-based services to that of a multi-dimensional service mode such as real-time virtual reference activities and video conferencing, etc. Needless to say, the real-time virtual reference modus operandi is much favored by library readers because of its novel features of timeliness and interactivity.

Real-time virtual reference services can be classified into 2 categories: 1) The people-based mode and 2) the artificial intelligence based mode. At present,
The people-based mode is much more prevalent. Generally speaking, this kind of real-time virtual service utilizes two types of tools: One is the tools specially designed for libraries. These tools, which include but not limited to Question Point\(^1\) (also called the 2nd generation of CDRS) and 24/7 Reference\(^2\), are commonly used in major overseas libraries as well as used by a few major Chinese libraries including such systems as the real-time reference service system\(^3\) of National Science and Technology Library, Collaborative Digital Service System\(^4-6\) of National Science Library, Chinese Academy of Sciences, Collaborative Online Reference Services (CORS)\(^7\) of Shanghai Library, Distributed Collaborative Virtual Reference System (CVRS)\(^8\) of China Academic Library & Information System (CALIS), etc. There are also other types of tools, which are designed for commercial applications, such as those instant messaging (IM) tools, which include QQ\(^9\), MSN\(^10\), open source software meebo\(^11\), etc.

The adoption of those specially designed software for people-based virtual library reference services has made remarkable progress in libraries in recent years. For example, as of December 2010, Question Point has 1,975 active users\(^12\), which provides services to readers who are from 32 countries. Similarly, in January 2010, the number of libraries abroad using IM tools totaled 1,011. In China, 117 out of 535 libraries surveyed provided IM reference services\(^13\).

This people-based real-time reference operation needs reference librarians on duty for its execution. This necessitates not only added labor costs, but also overlooked opportunities for effective and efficient delivery of reference services by means of using different social networking websites synchronously. It is suggested that by adopting applicable artificial intelligence technologies into virtual reference service operation, libraries can save labor costs and extend the range of reference services to a much broader reading public by means of accessing social networking spaces. More importantly, it can provide a ubiquitous service accessible in different geographical locations at all time. There are ample reports about such library practice both at foreign and domestic libraries. For example, a few relatively well-known cases abroad may include Ask Jeeves\(^14\) and Artificial Librarian of OPAL (Online Personal Academic Librarian) project, etc\(^15\). In China, libraries such as Shanghai Jiao Tong University Library and Tsinghua University Library have placed real-time reference service robots in operation, respectively\(^16-17\). The robot of Shanghai Jiao Tong University provides a self-assisted library service by asking users to choose their desired service from a menu list. The robot Xiaotu\(^18\) at Tsinghua University is interestingly characterized by its unique capability of handling users’ requested reference service through mutual interactive communication in the Chinese natural language.
2 Methodologies

2.1 Targeted goals

The objective of the Xiaotu project was aimed to provide, based on current artificial intelligence technologies, real-time virtual reference services online for 24/7 coverage year round. It was hoped that when Xiaotu communicates with users, she should be able to answer the library-related questions proficiently. She should be able to move from one topic to another with ease in a continuum of intellectual dialog with the user, or to provide a workable learning method for the inquirer to find a solution from another credible source to the issue being discussed. It was also presumed that in the system design, Xiaotu’s operation should be fully interfaced with the Library’s other online systems such as OPAC and social networking websites so as to optimize its virtual reference service potential.

Once a consensus was reached about the underlying principles of this project, we immediately embarked on the system design, which was to be made capable of realizing those said project objectives. As a result, we paid special attention to having the following features incorporated into the system design. These features included the functionalities of the system that are in line with user’s information seeking habits. It should have an interfacing operational capability for the performance of interactive chatting similar to that of MSN or QQ, etc. Moreover, we also devised some of our own unique ways to resolve a few technical issues. For instance, a viable and large corpus is needed for Xiaotu to operate. We resolved this issue in the following manner. As Tsinghua University Library has provided reference services for a long period of time and has accumulated a significant amount of valuable records, the contents of these records would constitute a rich source of corpus upon which Xiaotu can operate in comfort and in confidence.

Another example is in the area of user authentication. It was our intention that users need not go through the process of registration for seeking a library service so as to protect their identity privacy. What the library service users need to do is just to log onto a talk-screen to a virtual reference librarian, which eliminates the users’ concern about their privacy protection matters in using artificial intelligence technologies. In an operational aspect, our objective was not to make the user feel that he/she had been talking to a robot but rather to a real librarian on duty online. In the aspect of formulating suitable servicing websites, we took the following matters into consideration. As many social networking websites have been developed in recent years and have achieved admirable results, Xiaotu should be made conspicuously visible and accessible for her information service on those social networking websites so as to enable the robot to play a more important reference service role.
2.2 Selection

To build an artificial intelligence technology-supported real-time virtual reference system, a simple way is to have it built on the foundation of a mature application system rather than to have it started from scratch. MSN talking robot, XIAOI talking robot[19] and A.L.I.C.E.[20] are some tested and proven examples for our inspiration and emulation, especially for their talking capability. Tsinghua University Library carefully evaluated and experimented with each of these application systems in order to come up with a suitable model for steering our system design. Our study results are briefly summarized as follows.

MSN talking robot requires users to log onto the system and to initiate socializing steps with other system users so as to become groups of friends in order to operate. In addition, it is unable to access URLs accurately that contain Chinese characteristics, and it would push advertisements upon the users’ screen. The webpage of XIAOI robot has limited functions, although logging onto its chatting system is not needed. Moreover, it does not have adequate service support. At last, the open source software A.L.I.C.E. was chosen unanimously as a more suitable model for our Library’s system design due to its several outstanding features as detailed below.

A.L.I.C.E. is considered as a pioneer among all artificial intelligence robots. It was created by Richard S. Wallace in 1995. Wallace also invented AIML[21] (Artificial Intelligence Mark-up Language) to enable people to input knowledge into A.L.I.C.E. A large number of programmers around the world contributed to the successful development of AIML and it has different language versions. This greatly facilitated the use of A.L.I.C.E. on a world-wide scope[22]. A.L.I.C.E. has won the Loebner Prize Contest three times in 2000, 2001 and 2004, respectively.

A.L.I.C.E. adopts AIML as the descriptive language. Its success can be attributed to the following three aspects.

• Firstly, the knowledge database in AIML format presentation has a strong potential for its enrichment and expansion. Moreover, it possesses abundant grammar-driven functions of reasoning. The AIML embodied in A.L.I.C.E. is able to create and share knowledge in a single document conveniently. It can also be uploaded with a bundle XML tags into a multitude of AIML documents together so as to construct an even more intelligent robot.

• Secondly, A.L.I.C.E. has an effective reasoning mechanism[23]. Based on a module matching and the installation of a set of regulations for executing a session of heuristic and interactive conversation, A.L.I.C.E. can be equipped with the capabilities of learning, reasoning, judging, information storing, context acquiring and so on. It can serve as an auxiliary library staff for those unanswered questions related to information searching.
Thirdly, as A.L.I.C.E. system has stored information items over 40,000 subject categories\cite{20}, it can answer 95% of all the incoming questions in English. This cataloged information in the database of A.L.I.C.E. can be downloaded from its website\cite{24}.

Fig. 1 shows the workflow of A.L.I.C.E. When initiated, the pre-defined and related information such as names and gender are read into the system. The AIML knowledge database is uploaded into the memory in a tree-shaped structure. When user information input is received, the system divides the information input first into individual sentences and then has them processed into standardized questions. This is done by means of a technique of synonym substitutions and a few other relevant methods. The ability of using logic reasoning to detect the sentence structure of questions is the core function of the A.L.I.C.E. system. It will match the normalized sentence-structure of questions with those of similar sentence patterns stored in the tree-shaped memory mechanism in the system. The system tries to find the optimal matching results and to have them processed by templates after the matching inquiry items are found. For example, to have those specially marked information items contained in templates go through a process of reversion to their original form or to have those with jumping marks in the knowledge tree go through a process for performing further reasoning. Finally, responding results from the system to the inquiry input will be returned to users. The system will be ready for another inquiry input\cite{23}.

To sum up, A.L.I.C.E. has a robust architecture and sufficient knowledge bases. It is an excellent English reference service system in its own right. It was based on
Real-time virtual reference service based on applicable artificial intelligence technologies: The début of the robot Xiaotu at Tsinghua University Library

Fei YAO et al.
Research Papers

such a premise that we, the system designers, paid special attention to having its capabilities such as self-propelled initiatives in Chinese character recognition, Chinese word segmentation, platform integration, and so on incorporated into our system design.

2.3 Architecture

In view of the foregoing mentioned conceptualization about the adoption of an appropriate theoretical framework and functional features of A.L.I.C.E., we decided to use Java for writing the core programs of Xiaotu and have it mounted on a Linux operated server system. Its architecture is shown in Fig. 2.

This system architecture includes three major functional components as explained below:

- Graphical user interface (GUI) layer. It is responsible for the interaction with users. Users can input topics that they are interested in the dialog box in talking mode and for inquiring about books or other library-related information.
- Data I/O control layer. It is mainly responsible for the transfer and pre-processing of data. The pre-processing module is used to filter stop words such
as a, the, what, how, but that generally do not add meaning to a search and banned words that incite illegal activities and subversive speeches, obscenity and pornography, violence, murder and arson. It is also used to classify the text input by users and then to have them transferred to corresponding processing engines waiting for Xiaotu to show users the results. Data I/O control layer realizes the functions of the web server based on the Java servlet and establishes Ajax service requests between the GUI layer and itself in order to carry out the dynamic interaction required by the chat application.

- People-oriented natural language answering engine layer. It is responsible for specific inquiries and answers, including help information, book search engine, Baidu Baike search engine, training system, corpus searching and AIML knowledge database matching, etc.

2.4 Functions

The smart talking robot Xiaotu has two versions: The website version[8] and the application program version[25], and the latter is packaged into Renren website[26] for enhanced service accessibility. It plays the role of a real-time virtual reference librarian. The appearance of Xiaotu is shown in Fig. 3.

At present, the main functions and features of Xiaotu include:

- Natural language talking in Chinese. Xiaotu can talk with you about many different topics, such as weather, news events and so on. The conversational contents can be interesting and funny.
- Expert answers in professional fields. Xiaotu is able to give answers to questions specially related to Tsinghua University professors and the Library.
Real-time virtual reference service based on applicable artificial intelligence technologies: The début of the robot Xiaotu at Tsinghua University Library

- Book searching in the OPAC system. Users can search in Tsinghua University Library’s OPAC system directly in the dialog box.
- Baidu Baike searching. The search engine helps readers to search information in Baidu Baike, a third-party social database.
- Self-directed learning. Users can teach Xiaotu to learn new knowledge. Xiaotu can master them immediately and put them in practical application.
- Recommendation of hot topics for discussion. A list of high-profile topics is made readily available by Xiaotu for users to select directly for entering into the discussion.
- Natural language talking in English. Based on over 40,000 pieces of cataloged information items stored in database, Xiaotu is capable of handling about 95% of the posted English inquiries.

3 Key technical issues

3.1 Corpus

Searching for an information item in a corpus is based on the capability of the search engine[27]. Corpora are stored in XML format. An example of the process for searching a piece of a question-answer styled dialog stored in them looks like this:

```xml
<?xml version="1.0" encoding="utf-8"?>
<TPI_Record>
<QUESTION><![CDATA[How are you?]]></QUESTION>
<ANSWER><![CDATA[Fine. Thank you]]></ANSWER>
</TPI_Record>
```

In order to deal with the different levels of quality of resources from different corpora, we had these various groups of resources prioritized into the system. Specifically, the corpus of reference records accumulated through all the past years was assigned the top-tier priority. The corpus of questioning-answering information resources that were added manually was given a second-tier priority. The corpus of information resources that came from curriculum-related instructional programs was placed on the third-tier priority. Of course, the number of the graded levels of prioritization is not just limited to 3 levels.

3.2 Chinese word segmentation

Chinese word segmentation is one of the key points to achieve Chinese natural language stemming from a piece of questioning-answering styled dialog. Unlike Western languages such as English, there is no space between each word in Chinese writing. Word segmentation is a prerequisite before having them indexed. In the
field of the natural language processing, the Chinese word segmentation is currently a high-profile research issue. At present, there are two popular methods being applied to dealing with this issue. One is the mechanical word segmentation and the other, the word segmentation based on statistics. The mechanical word segmentation is a method of finding the maximum match in a direct forwarding or reversing match based on a dictionary of phraseology. Word segmentation based on statistics generally depends on the hierarchically hidden Markov Model. After a careful study and comparing several open source word segmentation applications, we adopted the Chinese Lexical Analysis System (ICTCLAS)\cite{28} created by the Institute of Computing Technology of Chinese Academy of Sciences.

ICTCLAS word segmentation has a rate of precision up to 98.45\%. The size of its API is below 200 kB. The compressed size of data from various dictionaries is less than 3 M. It is considered as one of the best Chinese grammar analyzers. After Chinese word segmentation, we added a procedure for eliminating some of the high frequency stop words such as those words used for acclamations, possessive cases, connections and some other censored sensitive words.

3.3 Similarity matching

The level of intelligence of a talking robot depends on the degree and extent of its being able to make correct matches between a group of questions and a group of corresponding answers. For each inquiry-response record stored in the corpus, its characteristic vectors are calculated by the term frequency-inverse document frequency (tf-idf). Term frequency \((t_f, i)\) denotes the frequency of word \(i\) in document \(j\), inverse document frequency \(idf\) denotes the ratio of the total number of documents and the number of the documents containing word \(i\), and the tf-idf weight of document \(j\) and word \(i\) is \(tf_{i,j} \cdot idf\).

After the tf-idf characteristic vectors of the inquiry-response record in corpora are calculated, similarity matching is calculated based on the cosine vector space model by the following equation:

\[
\text{Similarity (query, doc)} = \cos (d_i, d_j) = \frac{\sum_{k=1}^{n} \omega_{i,k} \omega_{j,k}}{|d_i| \cdot |d_j|} \quad (1)
\]

wherein, \(d_i\) and \(d_j\) represent the corresponding characteristic vectors of the question and the corpus record, respectively, and \(\omega_{i,k}\) denotes the tf-idf weight of word \(k\) in document \(j\).

Once Chinese word segmentation of different-level corpora is done, corresponding indexes are built. After receiving the user input and finishing word segmentation, the similarity matching can be done in the indexed records, and the optimal results could be ready for final output. The overall workflow of Chinese corpora searching module is shown in Fig. 4.
3.4 Book search engine

Book search engine depends on the INNOPAC\(^{[29]}\) system at Tsinghua University Library. Once a user inputs such a book as “Oliver Twist”, “Oliver Twist” will be used by the engine as the key words for accessing the INNOPAC system in retrieving the first 10 search results and sending the related information, such as the title of the book, authors’ names, publishing company, publishing year and so on to the inquirer. It also provides a URL on the result list by which the inquirer can access the INNOPAC system directly for other related or desired information.

3.5 Baike, a third-party resource search engine

Baidu Baike search engine is aimed to utilize pertinent resources on the Internet in support of the inadequate resources in the present Chinese corpus of Xiaotu. It indicates the possibility of integrating the third-party resources into this talking robot. Thus, we established a “crawler” system to allow it to lift desired information from Baidu Baike\(^{[30]}\). When users use the Baidu Baike search function for searching, Xiaotu processes the user input at the back end and to delegate the information searching task to Baidu Baike. If a matched result is found, then Xiaotu will grasp related information and generate the abstract and link in a repackaged format of
response to the inquirer. The inquirer can browse the abstract directly in the dialog box of Xiaotu or access Baidu Baiku based on the link provided for further needed information.

3.6 Training system

The training module embedded in Xiaotu is established in an inquiry-answer format. Users can use Xiaotu for curricula-related class instructions by inputting the contents of lectures in a “Q: question - A: answer” format. For the knowledge accumulated in such undertakings, the training system will have them effectively integrated into a training corpus for the future use.

4 An analysis of the job performance of Xiaotu

4.1 Application effects

Xiaotu attracted great attention since she came into being. In the first two months, library visitors averaged 3,000 people per day and more than 50,000 per month. The number of using her for classroom training purposes exceeded 3,600 requests. On the Renren website, visitors were above 7,000 on a monthly basis and the largest number of visitors in one day was about 400. The statistics of the homepage of Tsinghua University Library showed an increase of 10%–15% at the time of Xiaotu’s début. Xiaotu became one of the most popular topics in various library forums and was also constantly in the limelight resulting from the wide publicity by several news media[31].

However, owing to an overworking load of Xiaotu on the one hand and the unsatisfactory method of algorithm performed in the initial stage on the other hand, the virtual server was often shut down. Therefore, Xiaotu was subsequently placed into another server with a much improved version of the algorithm for better performance. Unfortunately, a lot of important statistical information was inadvertently lost during this period of Xiaotu’s migration.

After several months of trial and error, the usage of Xiaotu became stabilized. At present, the monthly usage of the service is over 2,000 times and the number of visitors per day is about 130. The number of using Xiaotu for training of some subject-specific topics per month is about 1,000 requests. On the Renren website, its usage also declined perceptibly. The number of average visitors to this website is about 130 per month and fewer than 10 visitors per day. Now, as far as the users’ background is concerned, only about 10% of the Renren website users are affiliated with Tsinghua University. The rest of those users are from the society at large, which reflects a fact that library services are extended to the reading public beyond the university campus walls. From the aspect of user’s task executions for each website
Real-time virtual reference service based on applicable artificial intelligence technologies:
The début of the robot Xiaotu at Tsinghua University Library

access, the average number of task executions is below 10, but the highest number registered during a reference session is about 90 times of inquiry-answer interactive encounters. Compared with the human being involved real-time reference system of CVRS we used in the library, the rate of usage frequency of Xiaotu is not any lower. On the assumption that the corpus is to be continuously enlarged and enriched, Xiaotu is expected to be able to deliver reference services at a much more sophisticated level in the not too distant future.

4.2 Discussion

There are still several unsolved problems for the improved job performance of Xiaotu. Firstly, as a new online reference service mode, Xiaotu has to answer questions more accurately in order to be embraced enthusiastically by users in meeting their information needs. That calls for a much expanded knowledge database to be installed in Xiaotu. Although the AIML knowledge databases in a few major Western languages, such as English, French etc., are of considerable scale, a sufficient large AIML Chinese knowledge database has yet to be developed. Up to now, Xiaotu has been equipped with an AIML knowledge database in Chinese of just a little over 8,000 subject items. Most of them are obtained from accumulated classroom lecturing materials with added manual editing. Be that as it may, it is still very difficult to achieve the requirement of being an intelligent robot in the strict sense.

Secondly, from the technological perspective, A.L.I.C.E. was originally developed for the use of Western language composed words and was not fully aware of the technical barriers for executing the Chinese word segmentation. The unique characteristics of the Chinese character structure made our development of an AIML knowledge database in Chinese very difficult. Moreover, in order to improve the quality level of intelligence of Xiaotu, complicated algorithm formulation was needed.

Thirdly, because Xiaotu is supposed to face all kinds of information seekers and provides services accordingly to numerous anonymous users, it often runs into some unconsidered people who would input derogatory language terms into Xiaotu for fun, which seriously affects the normal effective operation of Xiaotu as a result. Such abusive act has been solved only partially by adding some shielding words and the installation of an auditing mechanism into the system.

As from the perspective of practical usage, the non-academic related functions, such as poking funs with Xiaotu or engaging in self-directed reading in recreational literature, etc. have unfortunately attracted the most interest of library visitors. In contrast, reference services and searching functions are seldom judiciously invoked. This phenomenon, however, was not entirely unanticipated. It was thought
conceivable that library users needed a transitional period for more intensive guidance to access a library’s novel innovation. Above all, regardless any potential mishaps that may occur, we need continuously and tenaciously adopt more technology-assisted innovative methods in rendering professional information service par excellence.

4.3 Improvement

In view of the above-mentioned facts and defects, we need to develop, in a more conscientious way, not only a collection of enlarged library resources and related effective and efficient servicing mechanism but also a deep-rooted sense of pursuing professional excellence among library practitioners. We need to establish a mechanism so as to place the library mission, operating philosophy, practice and librarians’ professional growth under a systematic monitoring and review process in order to cope with a rapidly changing information environment. Specifically in our case at Tsinghua University Library, we need to establish a greatly enlarged and viable corpus of AIML knowledge database proactively. We also need to improve the storage, organization and retrieval of information resources more skillfully, so as to facilitate our mission-oriented tasks for knowledge sharing and knowledge creation. We, librarians at Tsinghua, are all fully committed to such undertakings and are working hard for their speedy materialization.

5 Conclusions

Information and library professionals have generally regarded that the adoption of applicable artificial intelligence technologies into real-time virtual reference services is a promising new online reference service *modus operandi*. It was with such a conviction and also based on the open source software A.L.I.C.E. and an accumulated rich experience in reference services, Tsinghua University Library has embarked on and accomplished much in instigating a virtual reference service project by capitalizing on its existing tangible and intangible assets for the design, development and deployment of an intelligent talking robot, named Xiaotu.

Xiaotu has quickly made her presence on the social networking website, named Renren website. In this way, Xiaotu has maximized her technological potential for the delivery of virtual library reference services. In the process of constructing Xiaotu, numerous new and existing applicable artificial intelligence technologies are incorporated into this venture. In conjunction with this undertaking, Tsinghua University Library also made a thorough inventory about its collection resources in order to forge concerted efforts for overhauling the Library’s operation in general and for the development of the Xiaotu project in particular.
A few bright spots in the development of Xiaotu may be worthy of mentioning. We abandoned, at a very early stage about the idea of having the project started from ground zero. Instead, we began it by making good use of mature open source software and applicable artificial intelligence technologies. In addition, we paid a specially focused attention on the feasibility of conducting virtual reference service operation in using the Chinese natural language. That was a pioneering accomplishment on our part among all libraries in China. It is our hope that our experience gained from working with this Xiaotu project would be a source of inspiration and a source of useful reference to those library colleagues who aspire to develop a similar type of real-time virtual reference operation at their own libraries.

References


(Copy editor: Ms. Lin PENG, Language revision: Prof. Charles C. YEN)